

## The Art of the Algorithm: Machine Learning in Environmental Health Research, with Nicole Kleinstreuer

Ashley Ahearn

We live in a time when investigators have overwhelming amounts of health-related data at their fingertips. In this podcast, Nicole Kleinstreuer explains how environmental health scientists are using machine learning to make sense of the information in those data—for example, predicting toxicological end points based on large curated data sets. But even as machine learning advances, researchers are working to set realistic expectations and performance thresholds for these new methods. <https://doi.org/10.1289/EHP6874>

NARRATOR [00:00:00] *EHP* presents “The Researcher’s Perspective.”

[Theme music]

AHEARN [00:00:09] It’s “The Researcher’s Perspective.” I’m Ashley Ahearn.

We live in a time where overwhelming amounts of health-related data are at our fingertips. Devices collect data on our heart rates and blood volume, the quality of the air we breathe, the contaminants that might be in the water we drink. We have broader population-level data and more precise satellite imagery about the world we live in than ever before. It’s a lot to process.

Artificial intelligence could help with that. In fact, AI and machine learning are already helping humans sift through all this data in order to better understand the ways in which our environment affects our health.<sup>1</sup>

Joining me to talk about the promise and potential pitfalls of artificial intelligence as it relates to environmental health research is Dr. Nicole Kleinstreuer. She’s the acting director of the Interagency Center for the Evaluation of Alternative Toxicological Methods within the National Toxicology Program at NIEHS. She’s also an associate editor for *Environmental Health Perspectives*. Dr. Kleinstreuer, great to have you on the show.

KLEINSTREUER [00:01:08] Thank you, Ashley. It’s wonderful to be here.

AHEARN [00:01:10] Could you give me some examples of how artificial intelligence and machine learning is already being applied to environmental health research?

KLEINSTREUER [00:01:18] Sure, absolutely. From my own experience and some of the work that we’re doing in our group, we have used machine learning and artificial intelligence to predict toxicological end points based on large curated data sets. So, if you have enough training data, you can really effectively leverage machine-learning algorithms to build predictive models that you have a high degree of scientific confidence in.

So we’ve looked at acute systemic toxicity, for example. So, this is an animal test that essentially looks at how poisonous a chemical might be—so, how much of a chemical do you have to accidentally ingest before it might kill you? And so, this is a test that’s typically run on rodents, so they feed a group of rats a particular chemical, and they increase the dose of that chemical until half the rats die, and then that’s called the lethal dose at which you see 50% mortality, or the LD<sub>50</sub>.

And that particular type of test has been run on tens of thousands of chemicals, and so, there are very large databases that we have curated where we’ve gotten training data. So, we’ve gotten those LD<sub>50</sub> values on over 15,000 chemicals, and those chemicals cover a very diverse molecular space. So there are many different types of molecules in that large training database. So, because we have such good coverage of the chemical space there, we can actually build models that just relate chemical structure—so the molecular features and the physical chemical properties of individual chemicals—to how toxic they might be. So it relates their

chemical structure to their acute oral systemic toxicity for over 15,000 chemicals.

And so, then we’ve worked with all these different computational modeling groups across the world to build what are known as quantitative structure–activity relationship models—so, QSAR models—that look at a structure of a chemical and predict how toxic it’s going to be, and all of that relies on machine-learning approaches. And so, we take a set of training data that’s diverse enough and represents enough different types of chemicals, and then we use that to train machine-learning algorithms to recognize what types of chemical structural features or physicochemical properties are associated with being toxic, are associated with being nontoxic. And then those predictive models can be used to predict the toxicity of hundreds of thousands of chemicals that have never been tested in these types of animal tests.

And to evaluate those types of predictive models, we hold out some of the existing data. And so, we use some of the data to train the models, usually about 80% of the data that we have in hand, and then we keep 20% of it back, and we don’t let the machine-learning algorithms “see” that 20%—that test data set—and we use it to evaluate the performance of those predictive algorithms.

So, those predictive algorithms are applied to chemicals that they’ve never seen before, but that we know the answers for, and then we can evaluate how well those predictive models are doing. And as it turns out, they are every bit as good as the animal test at predicting that toxicity end point as long as they’ve been trained on sufficient numbers of chemicals.

AHEARN [00:05:11] I’m really excited about the applications for public health research here. You know, scientists are using AI to develop predictive models that warn of toxic algal blooms<sup>2</sup> and to estimate air pollution levels.<sup>3,4</sup> I mean, that’s really exciting stuff that could protect a lot of people.

KLEINSTREUER [00:05:26] Right, I totally agree with you. Some of the work that’s being done in exposure biomonitoring is really fascinating, so things like designing wearable sensors to inform on chemical body burden and using things like product information from large retailers to predict population exposure to various chemicals. Those are all projects that really rely on machine learning at their heart and have been shown to be really, really effective in predicting what types of chemicals, you know, not just Americans but global citizens are being exposed to based on the areas where they live, the places where they shop, the products that they use, the activities that they engage in.

And that’s really, really important information for environmental health because it really helps prioritize—as far as we’re concerned in the toxicology space—where we should be spending our energy. Because it’s really important to know what types of chemicals people are being exposed to when you want to think about, you know, how we should be studying chemicals and where we should be testing and where we should be, you know, applying predictive models as far as toxicity is concerned.

AHEARN [00:06:50] So Dr. Kleinstreuer, what areas of environmental health research do you think could benefit most or make the biggest advances with the help of AI?

KLEINSTREUER [00:06:58] I think, you know, AI really has the potential to create some really exciting cross-disciplinary bridges between, for example, toxicology—so, some of the work that we and others are doing in building predictive models for toxic end points—and other disciplines like molecular epidemiology—so, using molecular epidemiological data to identify things like biomarkers of disease, biomarkers of exposure, and helping to prioritize where we should be spending our energy in terms of predicting toxic end points, in terms of what chemicals to study.

So, I think that's an area, that cross-disciplinary communication, where AI could really help to inform where, you know, we could spend our energy and what would be most interesting for us to study. Also, of course, in the area of exposure monitoring and understanding, you know, the connections between environmental chemical exposure and disease hot spots—I think that's another area where AI is really starting to make some real inroads into trying to help us kind of tease apart this incredibly complex relationship between the thousands of chemicals that we're exposed to on a day-to-day basis and how those truly impact our health status and our environment.

AHEARN [00:08:37] So I've got to ask about the pitfalls here. Do you think this technology could be oversold as some have cautioned?

KLEINSTREUER [00:08:44] Certainly. I think people tend to just think about, like, "Oh, AI and machine learning is just going to solve all of our problems" without actually thinking about what is the basis for AI and machine learning? Where is the training data coming from? What is the quality of that data? And based on the quality and the variability inherent in that training data, what are appropriate thresholds for performance and appropriate expectations for applications of AI and machine learning? So, I think that's one aspect of setting realistic expectations will really help us understand what are the best practical applications and utility of this type of technology.

AHEARN [00:09:28] Are you ever afraid of the Pandora's box that AI represents? You know there are more than a few science fiction books and movies about the robots taking over. . .

KLEINSTREUER [00:09:38] I, I, I. . . [laughs] No, not really [laughs]. I think that if we were actually that close to the robots, you know, being able to take over, I mean, I think I would be delighted with that because that would mean that we were light-years ahead of where I think we actually are in the field. I think that this kind of goes back to a point that I started to make earlier about the difference between artificial intelligence and augmented intelligence, and more and more I think we're recognizing, as a field, that some of the most effective applications of machines in things like environmental health research or in, you know, the medical field as a perfect example are in helping humans do their jobs better. So not in replacing humans, but in actually supplementing human intelligence and human intuition and human effectiveness.

So, a really good example is in the field of radiology, right? So, looking at scans, you can train machine-learning algorithms to recognize abnormalities and look at histopath images, but pathologists are still really as effective, if not more effective. And what's proven to be most effective is the combination of machines and humans. And so, the combination of machine-learning algorithms and artificial intelligence with human intelligence has proven to perform better than either of those components—either the machine on their own or the humans on their own. And so to me, that's really the

most exciting frontier, is leveraging these types of computational approaches to actually augment human intelligence and help us to do our jobs better.

AHEARN [00:11:38] So Dr. Kleinstreuer, this field is progressing so rapidly, and I'm wondering if I call you up in 10 years, what do you hope you're telling me about how artificial intelligence is being used to help protect and improve human health?

KLEINSTREUER [00:11:53] Well, I think that the field of environmental health is sort of catching up a little bit at the moment. There's the field of medicine, personalized medicine,<sup>5</sup> diagnostics—that's really been leading the charge in application of artificial intelligence to human health initiatives. And so, I hope that environmental health research will really start to catch up. And so, I would hope that in 10 years I can tell you about how excited I am that AI is being supported and enabled by large cross-cutting efforts like the NIH Data Commons initiative,<sup>6</sup> for example, that allows us to integrate massive data sources—from exposure data, to personalized medicine data, to toxicology data, to clinical data—in a way that allows us to really pinpoint environmental influences that are affecting human health. And then really rapidly and efficiently recommend solutions and take action in a much more expedient manner than we can currently without really sacrificing any level of scientific confidence and just increasing our efficacy.

AHEARN [00:13:18] Dr. Kleinstreuer, thank you so much for joining me.

KLEINSTREUER [00:13:21] Well, thank you so much for having me. It was wonderful.

AHEARN [00:13:24] Dr. Nicole Kleinstreuer is the acting director of the Interagency Center for the Evaluation of Alternative Toxicological Methods within the National Toxicology Program at NIEHS. She's also an associate editor for *Environmental Health Perspectives*.

I'm Ashley Ahearn. Thanks so much for listening to "The Researcher's Perspective."

[Theme music]

*The views and opinions expressed in this podcast are solely those of our guest and do not necessarily reflect the views, opinions, or policies of Environmental Health Perspectives or the National Institute of Environmental Health Sciences.*

## References and Notes

1. National Academies of Sciences, Engineering, and Medicine. 2019. *Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and Decisions: Proceedings of a Workshop—in Brief*. Washington, DC: National Academies Press. <https://www.nap.edu/catalog/25520/leveraging-artificial-intelligence-and-machine-learning-to-advance-environmental-health-research-and-decisions> [accessed 3 February 2020].
2. Yi H-S, Park S, An K-G, Kwak K-C. 2018. Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. *Int J Environ Res Public Health* 15(7):1322, PMID: 30248912, <https://doi.org/10.3390/ijerph15102078>.
3. Hong KY, Pinheiro PO, Weichenthal S. 2019. Predicting global variations in outdoor PM<sub>2.5</sub> concentrations using satellite images and deep convolutional neural networks. *arXiv.org/abs/1906.03975* [eess.IV].
4. Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J. 2016. Assessing PM<sub>2.5</sub> exposures with high spatiotemporal resolution across the continental United States. *Environ Sci Technol* 50(9):4712–4721, PMID: 27023334, <https://doi.org/10.1021/acs.est.5b06121>.
5. In personalized medicine, a care provider chooses certain practices, precautions, interventions, and products for a given patient on the basis of that individual's genetic makeup.
6. National Institutes of Health. 2019. New Models of Data Stewardship—Data Commons Pilot. [Website.] Updated 13 May 2019. <https://commonfund.nih.gov/commons> [accessed 3 February 2020].